BAB III

METODOLOGI PENELITIAN

3.1 Objek Penelitian

Objek penelitian yang terdapat pada penelitian ini adalah harga rumah di Kabupaten Tangerang, yang akan di prediksi harganya berdasarkan spesifikasi yang tercantum pada salah satu situs jual beli rumah yang ada di Indonesia, yaitu rumah123.com. Sumber data yang digunakan dalam penelitian ini didapatkan melalui proses *web scraping* pada situs rumah123.com. Adapun data yang di-*scraping* adalah data harga rumah, lokasi rumah, kamar tidur, kamar mandi, luas tanah, luas bangunan, sertifikat, daya listrik, jumlah lantai, kondisi properti, *carport*, kamar tidur pembantu, dan kamar mandi pembantu.

3.2 Metode Penelitian

Metode penelitian yang digunakan dalam penelitian ini adalah metode penelitian kuantitatif dengan pendekatan *data mining* menggunakan *framework* CRISP-DM. Jenis metode penelitian kuantitatif digunakan karena objek penelitian dalam penelitian ini adalah harga rumah di wilayah Tangerang, yang merupakan variabel dengan tipe data numerik. Pemilihan *framework* CRISP-DM digunakan karena *framework* CRISP-DM menyediakan tahapan iteratif yang lengkap sehingga proses *data mining* dapat dilakukan dengan jelas dari awal sampai akhir [29]. Berikut ini adalah perbandingan antara metode CRISP-DM dengan metode KDD dan SEMMA yang terlihat pada tabel 3.1 di bawah.

Tabel 3.1 Perbandingan Metode *Data Mining*

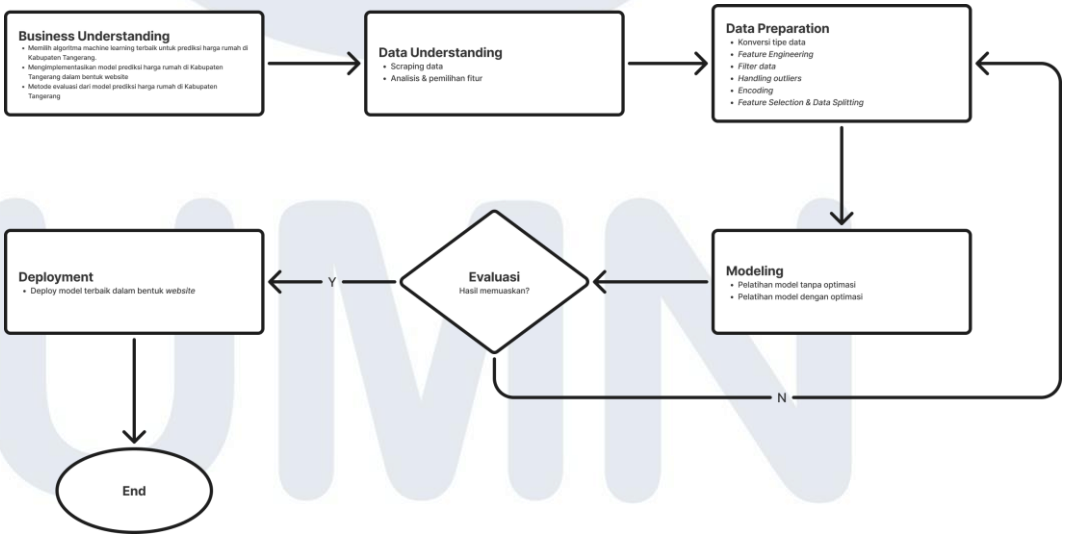
Aspek	CRISP-DM	SEMMA	KDD
Tahapan	*1. Business understanding* *2. Data understanding* *3. Data preparation* *4. Modeling* *5. Evaluation* *6. Deployment*	*1. Sample* *2. Explore* *3. Modify* *4. Model* *5. Assess*	*1. Selection* *2. Preprocessing* *3. Data transformation* *4. Data mining* *5. Interpretation evaluation*

Aspek	CRISP-DM	SEMMA	KDD
Fokus	Membahas alur dari pemahaman tentang masalah yang akan diselesaikan hingga proses *deployment*.	Hanya berfokus pada pemodelan data saja, tidak sampai *deployment.*	Pemahaman terhadap data kurang difokuskan pada KDD.

Sumber: [29], [43], [44]

3.2.1 Alur Penelitian

Dalam *framework* CRISP-DM, terdapat enam fase iteratif yang dimulai pada tahap *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment* [29]. Untuk rincian pada *framework* CRISP-DM, dapat dilihat pada ilustrasi 3.1 di bawah.



Gambar 3.1 Alur Metode CRISP-DM

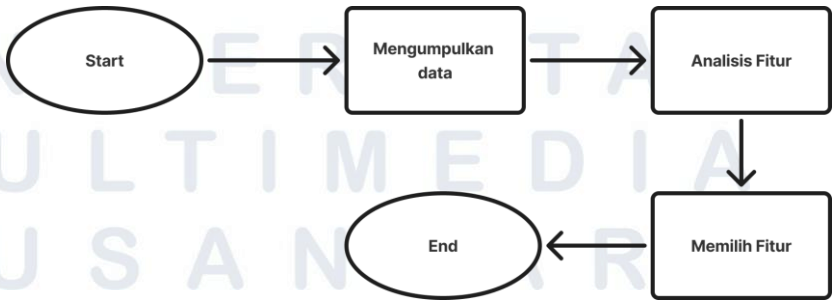
1. ***Business Understanding***

Tahap *business understanding* merupakan tahap pertama dalam metode *data mining* CRISP-DM. Dalam tahap ini, akan dilakukan analisa mengenai penentuan objek penelitian, tujuan penelitian, dan metrik keberhasilan dari penelitian yang dilakukan. Pada

penelitian ini, objek penelitiannya adalah data harga rumah di wilayah Kabupaten Tangerang, tujuan penelitian ini adalah untuk membuat model prediksi harga rumah sebagai harga acuan dalam membeli rumah di Kabupaten Tangerang, mengukur performa dari model prediksi yang telah dibuat (menggunakan nilai *mean absolute percentage error* (MAPE), *root mean squared error* (RMSE), dan R^2 (R-squared), dan melakukan *deployment* pada model *machine learning* terbaik dalam bentuk *website.*

2. *Data Understanding*

Tahap *data understanding* dilakukan untuk memahami data yang digunakan dalam penelitian. Pemahaman mengenai data diperlukan agar tujuan dari penelitian dapat tercapai. Adapun, tahapan yang dilakukan pada tahap *data understanding* pada penelitian ini dimulai dengan pengumpulan data. Data yang akan digunakan dalam penelitian ini dikumpulkan melalui teknik *web scraping* yang dilakukan pada situs rumah123.com. Hasil *web scraping* pada situs rumah123.com menghasilkan data mentah sebesar 7145 baris dan 17 kolom yang berisi data harga rumah yang terpasang di situs rumah123.com yang berlokasi di seluruh Tangerang Raya. Setelah itu, akan dilakukan proses *filter* data agar data yang digunakan dalam pemodelan nanti adalah data harga rumah yang berlokasi di Kabupaten Tangerang serta melakukan analisis pada *dataset* untuk menemukan apakah ada atau tidaknya anomali dari *dataset*. Rincian dari tahapan *data understanding* dapat dilihat pada *flow chart* 3.2 berikut.



Gambar 3.2 *Flowchart* Tahap Data *Understanding*

Selanjutnya, penjelasan dari nama variabel yang akan digunakan dapat dilihat pada tabel 3.2 di bawah.

Tabel 3.2 Variabel Data yang akan digunakan

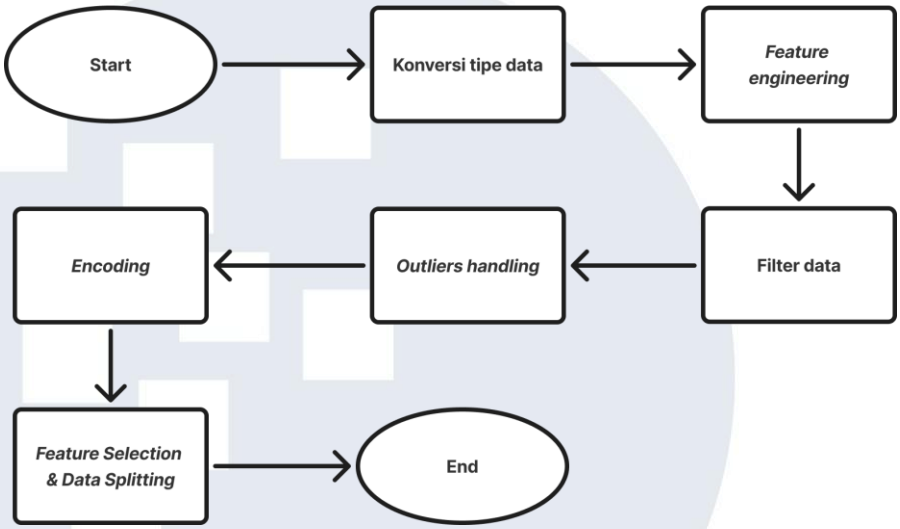
Kolom	Keterangan
URL	URL *Listing* properti di rumah123.com
Harga	Harga rumah (dalam bentuk *string*)
Lokasi	Lokasi rumah
Kamar Tidur	Jumlah kamar tidur
Kamar Mandi	Jumlah kamar mandi
Luas Tanah	Luas tanah properti
Luas Bangunan	Luas bangunan properti
Sertifikat	Sertifikat properti
Daya Listrik	Daya listrik dalam satuan Watt
Jumlah Lantai	Jumlah lantai yang dimiliki properti
Kondisi Properti	Kondisi properti (Baru, Bagus, Butuh renovasi)
Lainnya	Informasi yang gagal didapat ketika *scraping*
Carport	Jumlah mobil yang dapat masuk ke *carport*
Kamar Tidur Pembantu	Jumlah kamar tidur pembantu
Kamar Mandi Pembantu	Jumlah kamar mandi pembantu
Garasi	Jumlah mobil yang dapat masuk ke garasi
IMB	Izin mendirikan bangunan (Ada/Tidak)

Berdasarkan informasi pada tabel 3.2 di atas, kolom URL, Lainnya, Garasi, dan IMB tidak akan dipakai karena keempat kolom di atas memiliki nilai *null* dan tidak memiliki relevansi dengan model yang akan dibuat kedepannya.

3. *Data Preparation*

Pada tahap ini akan dilakukan proses konversi kolom ke tipe data yang tepat, melakukan proses filter pada *dataset*, *feature engineering*, *outliers handling*, dan melakukan *encoding* dengan *One-Hot Encoder* dan *OrdinalEncoder* pada kolom kategorikal. Konversi kolom ke dalam tipe data yang tepat dapat membantu model untuk memahami hubungan antara satu variabel dengan variabel lainnya. Kolom-kolom yang akan di konversikan ke dalam tipe data yang tepat antara lain kolom harga, jumlah kamar tidur, jumlah kamar mandi, luas tanah, luas bangunan, daya listrik, jumlah lantai, *carport*, kamar tidur pembantu, dan kamar mandi pembantu. *Feature engineering* akan dilakukan untuk mengekstrak nama kecamatan dari kolom Lokasi. Proses filter dilakukan untuk mengeliminasi data-data spesifikasi rumah yang tidak masuk akal yang terdapat pada kolom *carport*, sertifikat, dan kondisi properti. Selanjutnya, akan *outliers handling* dengan cara melakukan *log transformation* pada kolom Harga sehingga memiliki distribusi normal. Terakhir, *encoding* akan dilakukan sebanyak dua kali: Pertama, *encoding* menggunakan *OneHot-Encoder* pada tiap kolom kecamatan yang sudah dilakukan *feature engineering* sebelumnya. Kedua, *encoding* menggunakan *OrdinalEncoder* pada kolom sertifkat, kondisi properti, dan daya listrik. Setelah dilakukan proses persiapan data, dataset yang digunakan memiliki 34 kolom dan 1806 baris. Selanjutnya, akan dilakukan *feature selection* dan *data splitting* untuk mempersiapkan data sebelum dilatih ke dalam model *machine learning* yang akan dibuat.

Langkah-langkah yang dilakukan dapat dilihat pada gambar *flowchart* 3.3 di bawah.

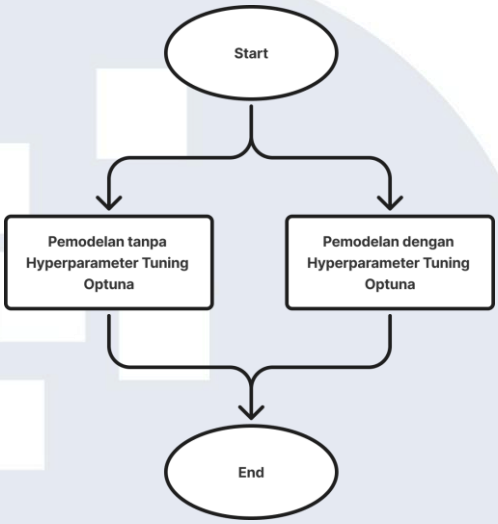


Gambar 3.3 *Flowchart* Tahap *Data Preparation*

4. *Modeling*

Setelah tahap persiapan data dilakukan, hasil luaran dari tahap persiapan data akan digunakan pada tahap pemodelan. Tahap pemodelan akan dilakukan dua kali, yaitu pemodelan pada algoritma *Multiple Linear Regression, Random Forest,* dan *XGBoost* tanpa *hyperparameter tuning* Optuna dan dengan *hyperparameter tuning* Optuna. Pemodelan dilakukan sebanyak dua kali untuk mendapatkan komparasi antara model yang belum di optimasi dengan model yang telah di optimasi. Kedua model akan divalidasi dengan *K-Fold cross validation* untuk menilai performa model apakah terindikasi *overfitting/underfitting* pada model yang telah dibuat.

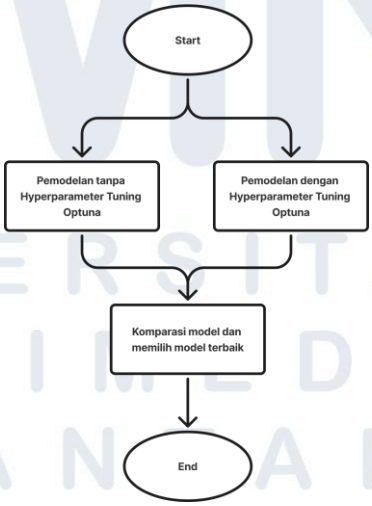
Rincian dari alur yang diterapkan pada tahap modeling dapat dilihat pada gambar *flowchart* 3.4 di bawah.



Gambar 3.4 *Flowchart* Tahap *Modeling*

5. *Evaluation*

Pada tahap evaluasi, akan dilakukan evaluasi performa dan komparasi model dari model yang telah dipilih dan dibangun pada tahap pemodelan sebelumnya. Komparasi model akan dilakukan berdasarkan performa model yang ditunjukkan dengan metrik evaluasi *R-squared,* MAPE, dan RMSE. Model *machine learning* yang terbaik adalah model yang menghasilkan perfroma *R-squared* yang tinggi (mendekati 1) dan nilai MAPE serta RMSE yang rendah.



Gambar 3.5 *Flowchart* Tahap *Evaluation*

6. ***Deployment***

Pada tahap terakhir dari metode *data mining CRISP-DM*, yaitu tahap *deployment*, akan dilakukan pembuatan *website* menggunakan *package streamlit* pada model prediksi harga rumah dengan akurasi terbaik. Pada *website* juga akan ditambahkan sebuah fitur untuk memprediksi spesifikasi rumah berdasarkan lokasi dan *budget* yang di-*input* oleh pengguna. *Website* yang dibuat akan ditujukan untuk pengguna yang sedang mencari rumah di area Kabupaten Tangerang sesuai dengan *budget* yang pengguna miliki.

3.3 Variabel Penelitian

3.3.1 Variabel Dependen

Variabel dependen adalah variabel yang nilainya dipengaruhi oleh variabel independen. Dalam penelitian ini, variabel harga bertindak sebagai variabel dependen. Variabel harga dipilih karena harga rumah bergantung pada spesifikasi yang rumah itu tawarkan.

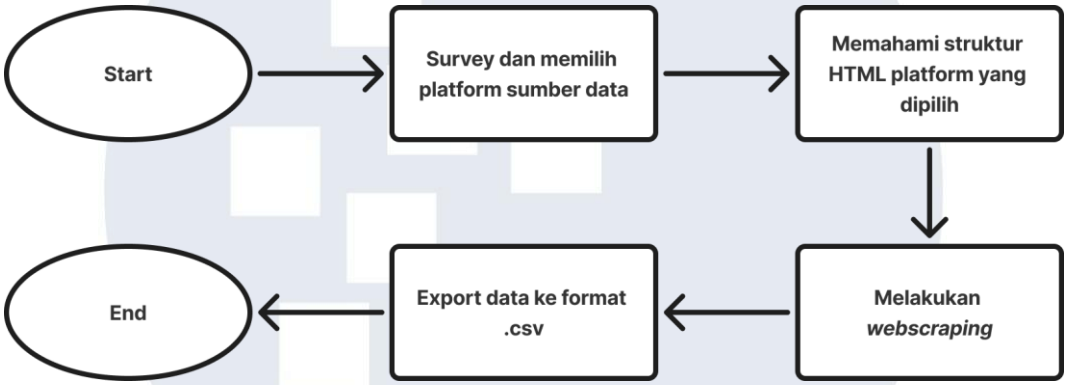
3.3.2 Variabel Independen

Variabel independen adalah variabel yang nilainya tidak bergantung pada variabel lain dan dapat memengaruhi nilai variabel dependen. Dalam penelitian ini, selain variabel harga merupakan variabel independen.

3.4 Teknik Pengumpulan Data

Tahapan yang dilakukan untuk pengumpulan data dimulai dengan survei situs jual beli properti yang ada di Indonesia sebagai sumber data penelitian. Setelah melakukan survei dan memilih situs jual beli properti, langkah selanjutnya adalah memahami struktur HTML *website* yang dipilih (rumah123.com) untuk memperoleh informasi-informasi yang akan digunakan, seperti tautan URL, informasi luas tanah dan luas bangunan rumah. Data-data URL dari tiap rumah yang tayang di situs rumah123.acom akan diperoleh menggunakan *package Selenium*, sedangkan data-data spesifikasi rumah (harga, lokasi, dll) akan diperoleh

menggunakan *package BeautifulSoup4* untuk melakukan *scraping* dari situs jual beli rumah yang dipilih. Selanjutnya, data yang telah sukses diperoleh akan disimpan ke dalam format *csv* untuk diolah lebih lanjut. Rincian dari tahapan yang digunakan pada teknik pengumpulan data ini dapat dilihat pada gambar *flowchart* 3.6 berikut.



Gambar 3.6 *Flowchart* Alur Teknik Pengumpulan Data

3.5 Teknik Analisis Data

Dalam penelitian ini, akan digunakan algoritma *machine learning* untuk membantu proses analisis data. Tabel 3.5.1 berisi perbandingan dari algoritma regresi linear, *random forest regressor, XGBoost, decision tree,* dan KNN [10] [11] [14] [45] [46]. Berdasarkan perbandingan yang telah dijelaskan pada tabel 3.5.1, algoritma *machine learning* yang dipilih dalam penelitian ini adalah algoritma *Multiple Linear Regression, Random Forest,* dan *XGBoost.* Ketiga algoritma tersebut dipilih karena berdasarkan hasil SLR yang telah dilakukan, ketiga algoritma tersebut banyak digunakan dalam kasus prediksi harga rumah. Selain itu, alasan pemilihan algoritma *Multiple Linear Regression* adalah sebagai *base model* atau patokan ketika melakukan komparasi, algoritma *Random Forest* dan *XGBoost* digunakan karena kedua algoritma ini cenderung memiliki performa yang lebih baik dan tahan terhadap *overfitting.*

Tabel 3.3 Perbandingan Kelebihan & Kekurangan Algoritma

Algoritma	Kelebihan	Kelemahan
Regresi linear	• Mudah dipahami dan diimplementasi. • Dapat mengetahui hubungan antara dua variabel.	• Rawan terjadi *overfitting & underfitting.* • Sensitif terhadap *outliers.*
Random forest	• Kuat terhadap *outliers.* • Risiko terjadinya *overfitting* yang rendah. • Bekerja dengan efisien pada *dataset* dengan ukuran yang besar	• Waktu pelatihannya lambat. • Membutuhkan sumber daya komputasi yang lebih besar.
XGBoost	• Akurasi yang tinggi. • Cepat dan efisien. • Dapat digunakan pada tugas klasifikasi atau regresi.	• Algoritma yang kompleks. • Memerlukan sumber daya komputasi yang besar untuk mengolah data yang besar.
Decision Tree	• Dapat digunakan untuk kasus klasifikasi/regresi. • Mudah di interpretasikan. • Tahan terhadap *outliers.*	• Rawan terjadi *overfitting.* • Tidak konsisten apabila terdapat perubahan data.
K-NN	• Dapat digunakan untuk kasus klasifikasi/regresi. • Mudah diimplementasikan. • Modelnya konsisten apabila terdapat perubahan/penambahan data.	• Sensitif dengan *noise* atau *missing values.* • Performa kurang optimal dengan *dataset* dengan ukuran yang besar dan dimensionalitas yang tinggi.

Sumber: [10] [11] [14] [45] [46]